1. Montgomery Building Permits – Classifying Permit Types   
   Complexity: Low/Med
   1. Business Understanding: City employees must manually enter the classification of each building permit into a computer system. A program that would automatically classify the “Use Type” as “Commercial” or “Residential” based on data already being entered by the user (e.g. address, business name, project synopsis) could lead to fewer errors of classifications and allow employees to concentrate on other tasks saving the city time and money.
   2. Data understanding: Building permit data is publicly available via a Socrata website and contains records from 2014 to present. There are 17.5K rows, and 35 columns. The url of the data is <https://data.montgomeryal.gov/Permits/Building-Permit-2014-Present-Download-/qvzc-ejq2>

First run summary statistics, histograms, correlation coefficients, and other methods to identify key features to align data understanding with the business understanding.

* 1. Data Preparation: File is generally clean but likely includes missing values, outliers, and unnecessary fields. Identify key features for the model. Natural Language Processing may be employed to help with classifications.
  2. Modeling: Potential candidates for modelling include Logistic Regression, Decision Tree, or Support Vector Machines. May also consider a multi-label classifier for “permit type” instead.

1. San Francisco Fire Calls for Service – Predicting Emergency Levels  
   Complexity: Med/High
   1. Business Understanding: The city has a large call for service volume, but most calls are ultimately not life-threatening. However, those incidents that were serious included public and/or rescuer fatalities, injuries, and/or property damage. If first responders had an accurate rating system that predicted the level of emergency based on the incoming 911 call data, 911 operators and/or fire chiefs may choose different levels of personnel and/or apparatus to dispatch to the call location.
   2. Data understanding: The data is publicly available via Socrata’s open data platform. There are 4.64M rows, 34 columns with call records from 2000 to present.

<https://data.sfgov.org/Public-Safety/Fire-Department-Calls-for-Service/nuek-vuh3>

Run summary statistics, histograms, correlation coefficients, and other methods to identify key features. A clustering algorithm may be helpful to identify class candidates. Must determine if there are enough records for each classification to make accurate predictions.

* 1. Data Preparation: Data needs to be classified based on multiple features. Number of public and/or rescuer fatalities and/or damage costs could be used to create three labels of emergency classification (high, med, low). Missing values and outliers may need to be dropped. Standardization and/or dimensionality reduction should also be considered to reduce model error/effects of outliers and computational costs.
  2. Modeling: Potential candidates for a multi-class model include Linear Discriminant Analysis, Naïve Bayes or K-Nearest Neighbors. Standardization and Principal Component Analysis should be considered. Caution: The rather large volume of data under consideration may lead to moderately high computational costs. In addition, there could be too few records with a “high” classification, for example, to accurately make predictions for high level emergencies. The error rate for such modelling would need to be investigated using different algorithms.

1. GAIA Telescope Observations – Identifying Threats  
   Complexity: Vert High
   1. Business Understanding: In addition to star data, the GAIA observatory recorded many objects within the solar system that could be potential dangers to Earth. Using regression, determine which objects are heading toward earth the fastest, that may pose a threat, or that have particular interesting characteristics yet to be determined.
   2. Data understanding: The object data contains a sample of 14,099 SSOs for a total of total of 1,977,702 different observations. The data can be found at <http://cdn.gea.esac.esa.int/Gaia/gdr2/>

Summary statistics, histograms, correlation coefficients and other methods will be useful to determine key features. Dimensionality reduction or clustering may be employed to identify classifications of objects.

* 1. Data Preparation: SSO observations are contained in four separate CSV files which will need importing and joining. The complexity of the data and features are unknown at this time and could be overwhelming to the novice. Sampling down the number of observations may be necessary to reduce computational time. Standardization and Principal Component Analysis are almost certainly necessary as many fields are numeric.
  2. Modeling: Potential candidates for modelling may include Support Vector Machines, K-Nearest Neighbors, Random Forests or even Neural Networks. Caution: The very large volume of data under consideration may lead to computational costs where a cloud computing environment might be necessary.